quiz5_210615

2024-11-06

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

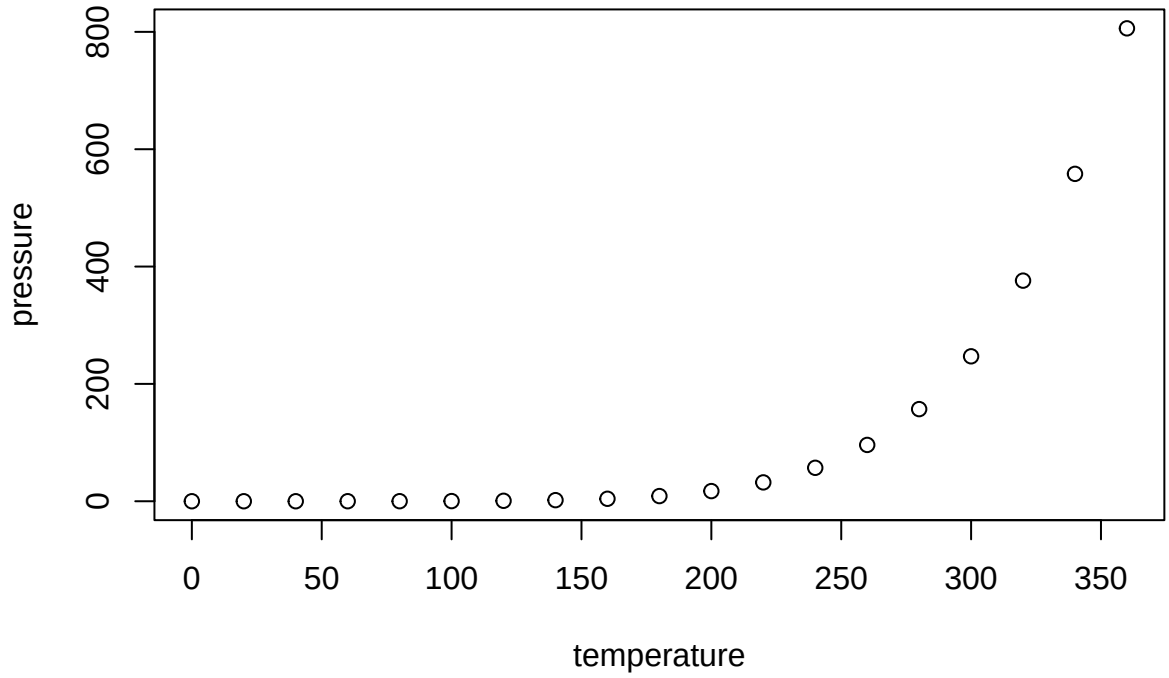
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
data("airquality")  # Load the dataset
airquality <- na.omit(airquality)  # Remove rows with NA values
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 7    23     299  8.6   65     5   7
## 8    19      99 13.8   59     5   8
```

```
lm_simple <- lm(Ozone ~ Temp, data = airquality)
summary(lm_simple)
```

```
##
## Call:
## lm(formula = Ozone ~ Temp, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.922 -17.459  -0.874  10.444 118.078
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.6461    18.7553  -7.872 2.76e-12 ***
## Temp           2.4391     0.2393  10.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.92 on 109 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4833
## F-statistic: 103.9 on 1 and 109 DF,  p-value: < 2.2e-16
```

The p value is here is very less stating that we can reject the null hypothesis of temp not affecting the ozone and slope is positive inferring a positive correlation

```
lm_null <- lm(Ozone ~ 1, data = airquality)  # Start with an empty model
lm_full <- lm(Ozone ~ Temp + Solar.R + Wind + Month, data = airquality)  # Full model with all predicto
forward_model <- step(lm_null, scope = list(lower = lm_null, upper = lm_full), direction = "forward")
```

```
## Start:  AIC=779.07
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     59434  62367 706.77
## + Wind     1     45694  76108 728.87
## + Solar.R  1     14780 107022 766.71
## + Month    1      2487 119315 778.78
## <none>                 121802 779.07
##
## Step:  AIC=706.77
## Ozone ~ Temp
##
##           Df Sum of Sq   RSS    AIC
## + Wind     1   11378.5 50989 686.41
## + Month    1    2824.7 59543 703.63
## + Solar.R  1    2723.1 59644 703.82
## <none>                 62367 706.77
##
## Step:  AIC=686.41
## Ozone ~ Temp + Wind
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2986.2 48003 681.71
## + Month    1    2734.8 48254 682.29
## <none>                 50989 686.41
##
## Step:  AIC=681.71
## Ozone ~ Temp + Wind + Solar.R
##
##         Df Sum of Sq   RSS    AIC
## + Month  1    1701.2 46302 679.71
## <none>               48003 681.71
##
```

```
## Step:  AIC=679.71
## Ozone ~ Temp + Wind + Solar.R + Month
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = Ozone ~ Temp + Wind + Solar.R + Month, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.870 -13.968  -2.671   9.553  97.918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -58.05384   22.97114  -2.527   0.0130 *
## Temp          1.87087    0.27363   6.837 5.34e-10 ***
## Wind         -3.31651    0.64579  -5.136 1.29e-06 ***
## Solar.R       0.04960    0.02346   2.114   0.0368 *
## Month        -2.99163    1.51592  -1.973   0.0510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 106 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.6055
## F-statistic: 43.21 on 4 and 106 DF,  p-value: < 2.2e-16
```

the forward model selection method helped in sequentially selecting the varibales on theie AIC values . Variables like temp, wind, solar radiation had with great confidence value (low p values <0.05 suggest it) but the the model also included month but it did have a signicifant p-value and thus not so confidence so it shows it might have a weaker effect.

```
backward_model <- step(lm_full, direction = "backward")
```

```
## Start:  AIC=679.71
## Ozone ~ Temp + Solar.R + Wind + Month
##
##           Df Sum of Sq   RSS    AIC
## <none>                 46302 679.71
## - Month    1    1701.2 48003 681.71
## - Solar.R  1    1952.6 48254 682.29
## - Wind     1   11520.5 57822 702.37
## - Temp     1   20419.5 66721 718.26
```

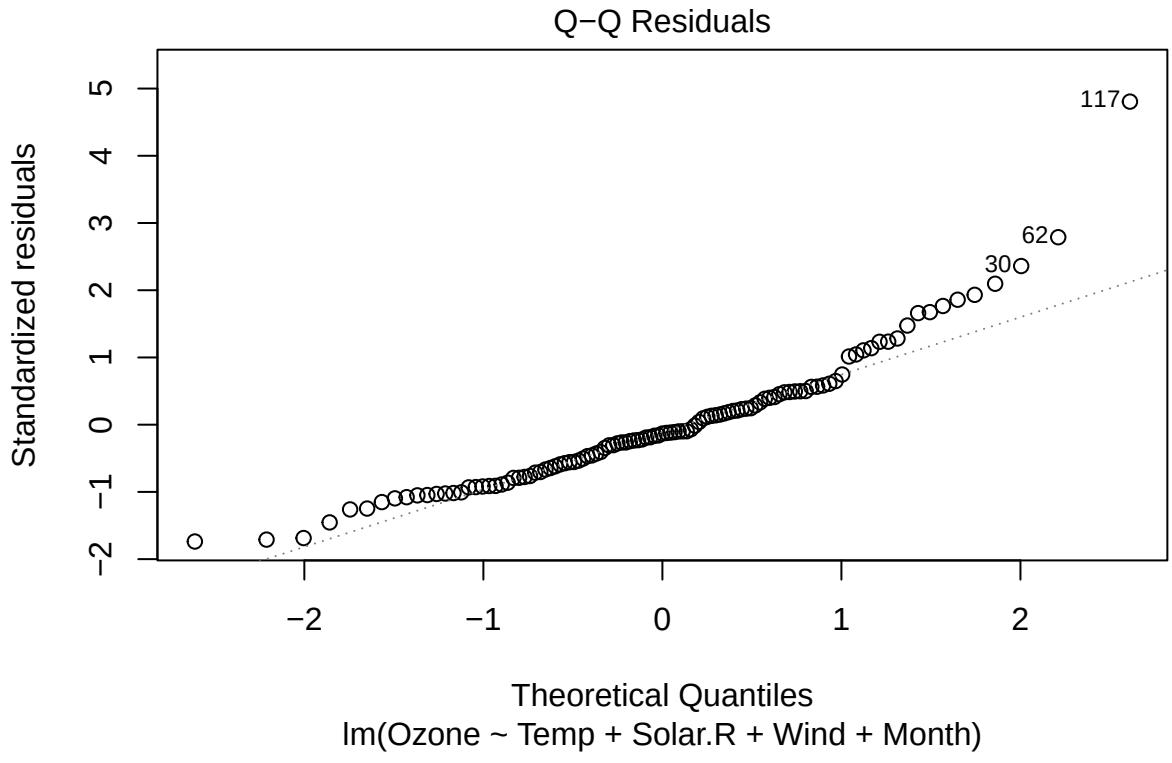
```
summary(backward_model)
```

```
##
## Call:
## lm(formula = Ozone ~ Temp + Solar.R + Wind + Month, data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -35.870 -13.968  -2.671   9.553  97.918
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -58.05384   22.97114  -2.527   0.0130 *
## Temp          1.87087    0.27363   6.837 5.34e-10 ***
## Solar.R       0.04960    0.02346   2.114   0.0368 *
## Wind         -3.31651    0.64579  -5.136 1.29e-06 ***
## Month        -2.99163    1.51592  -1.973   0.0510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 106 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.6055
## F-statistic: 43.21 on 4 and 106 DF,  p-value: < 2.2e-16
```

The backward selction also showed similar result as forward model but with increase oreder for selction oF AIC score . The R-squared value is 0.6199, indicating that approximately 61.99% of the variance in Ozone levels is explained by this model. The F-statistic in both also has significant p-value showing it has strong model for prediction.

```
plot(x=lm_full, which = 2)  # QQ plot for residuals
```




```
shapiro.test(residuals(lm_full))  # Shapiro-Wilk test for normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_full)
## W = 0.91646, p-value = 3.341e-06
```

The test fails normaility

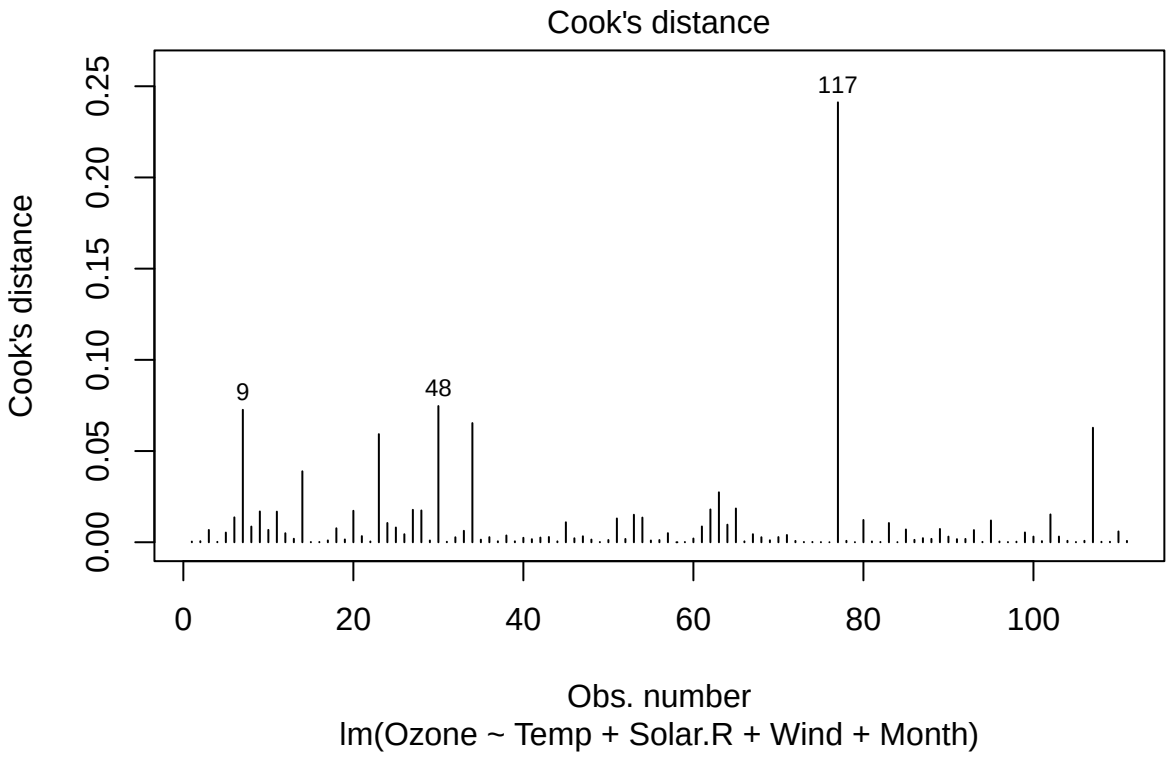
```
residuals( object = lm_full )
```

```
##           1           2           3           4           7           8
##   3.7822070  -5.0112848 -19.0345907  -2.3661453 -11.9022883  22.4881687
##           9          12          13          14          15          16
##  32.6081765 -20.6049596 -23.3368850 -17.6470174  23.0554289 -11.1494262
##          17          18          19          20          21          22
##   8.1061952  29.5273929  -2.0377595  -1.9943157  -4.5962086 -13.3787368
##          23          24          28          29          30          31
##  -6.1811022  26.1238820   9.8167886   3.3887912  48.0569603 -21.4697747
##          38          40          41          44          47          48
## -22.5367002   9.9601184 -25.6423303 -35.2162996  -7.1106794  32.8669392
##          49          50          51          62          63          64
##   3.0736138 -16.3819415 -25.8175202  57.0979780 -12.8171645 -21.7385074
##          66          67          68          69          70          71
##  -5.7107918 -15.7107415  -5.4161856  11.5265020   9.2886122  13.3501918
##          73          74          76          77          78          79
## -13.2460304  -4.8089974 -18.6292322 -14.5568029 -18.8459003 -10.3992506
##          80          81          82          85          86          87
##   2.8688058  10.1995183 -20.9126892  12.0406020  43.4429449 -29.9117787
##          88          89          90          91          92          93
##   5.8312613  10.3363862 -20.9968695  -0.2931180   4.3687497 -11.7865367
##          94          95          99         100         101         104
## -15.9764090 -34.7015748  38.0979415  25.4105945  39.8737531  -6.2910165
##         105         106         108         109         110         111
## -18.8248643  21.7003982  -9.4317476  11.5524037 -18.3610115  -8.8929669
##         112         113         114         116         117         118
##   4.7953562  -2.5101022   1.9245478   0.8434465  97.9181722   9.9604736
##         120         121         122         123         124         125
##  -1.3859102  20.5934114  -4.4774605   2.6945316  25.3302199  -1.9982757
##         126         127         128         129         130         131
## -15.8027655   7.8693705 -10.9570587   6.6680792 -21.0398612 -14.7009208
##         132         133         134         135         136         137
##  -9.5943621  -8.2707338  15.1488519   2.3523994 -21.9888275  -3.8939206
##         138         139         140         141         142         143
##  -2.2685357  -3.8201988  12.2880864 -11.3869856   4.1150742 -35.8700478
##         144         145         146         147         148         149
##   8.2265409   4.9639817  -3.2961992  -5.3819993  35.1755606  -2.6709482
##         151         152         153
##  -3.3839538 -19.1730272   4.8388379
```

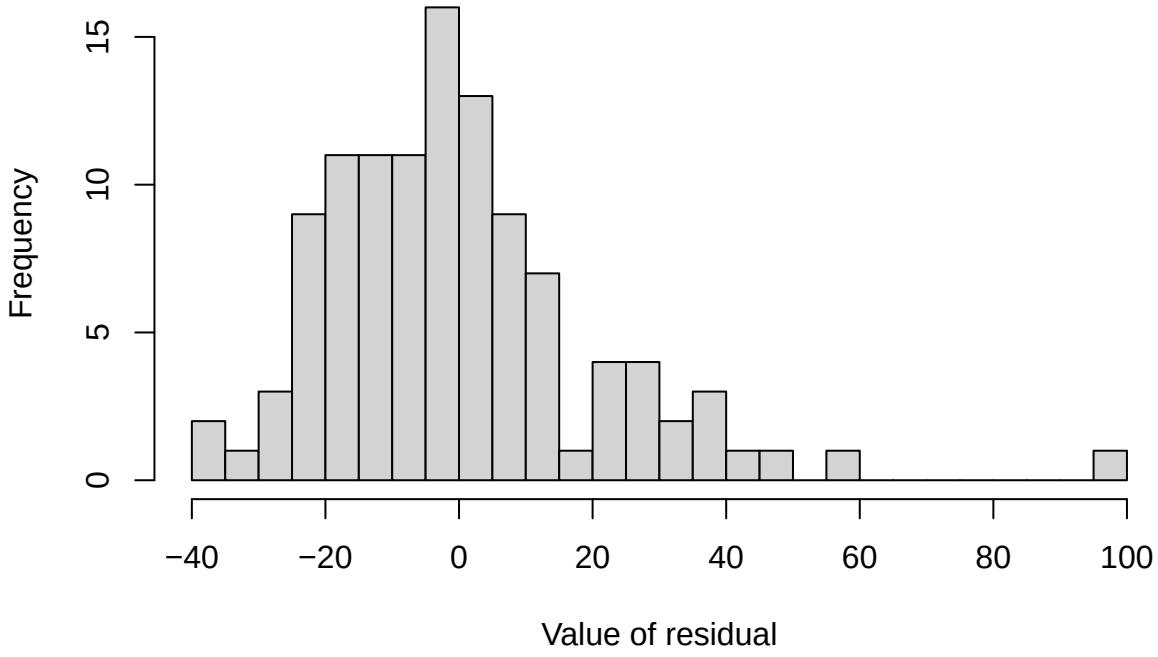
```
#checking for the outliers using cooks distance
cks <- cooks.distance( model = lm_full)
cks
```

```
##            1            2            3            4            7            8
## 3.819610e-04 5.689341e-04 6.737908e-03 2.210400e-04 5.297461e-03 1.364310e-02
##            9           12           13           14           15           16
## 7.259880e-02 8.593804e-03 1.688362e-02 6.759613e-03 1.677788e-02 4.933286e-03
##           17           18           19           20           21           22
## 1.914584e-03 3.890207e-02 1.217570e-04 1.395052e-04 9.998281e-04 7.654326e-03
##           23           24           28           29           30           31
## 1.532209e-03 1.725082e-02 3.352739e-03 4.232986e-04 5.926914e-02 1.055027e-02
##           38           40           41           44           47           48
## 8.072954e-03 4.400754e-03 1.777012e-02 1.742435e-02 9.601515e-04 7.466466e-02
##           49           50           51           62           63           64
## 2.519029e-04 2.728151e-03 6.273281e-03 6.532376e-02 1.416770e-03 2.793499e-03
##           66           67           68           69           70           71
## 5.171530e-04 3.703762e-03 5.005981e-04 2.438342e-03 1.669900e-03 2.570028e-03
##           73           74           76           77           78           79
## 2.868304e-03 4.779540e-04 1.094586e-02 2.168611e-03 3.266233e-03 1.455897e-03
##           80           81           82           85           86           87
## 1.254744e-04 1.227731e-03 1.303843e-02 1.792987e-03 1.507375e-02 1.352208e-02
##           88           89           90           91           92           93
## 9.593055e-04 1.213779e-03 4.979836e-03 7.467969e-07 1.318007e-04 2.049014e-03
##           94           95           99          100          101          104
## 8.647946e-03 1.797691e-02 2.738321e-02 9.624625e-03 1.851116e-02 4.849019e-04
##          105          106          108          109          110          111
## 4.416633e-03 2.738917e-03 1.077415e-03 2.830742e-03 3.966377e-03 7.108318e-04
##          112          113          114          116          117          118
## 1.341255e-04 1.575414e-04 8.730484e-05 4.386719e-06 2.411814e-01 7.723968e-04
##          120          121          122          123          124          125
## 5.954998e-05 1.220949e-02 4.298462e-04 1.384957e-04 1.054343e-02 7.541471e-05
##          126          127          128          129          130          131
## 6.992822e-03 1.315716e-03 2.229412e-03 1.830503e-03 7.281488e-03 3.078091e-03
##          132          133          134          135          136          137
## 1.698170e-03 1.809277e-03 6.655904e-03 1.927503e-04 1.192839e-02 4.134921e-04
##          138          139          140          141          142          143
## 9.873540e-05 3.006155e-04 5.365458e-03 3.088692e-03 5.907506e-04 1.523985e-02
##          144          145          146          147          148          149
## 3.113280e-03 7.628859e-04 1.295501e-04 7.851032e-04 6.276483e-02 2.465907e-04
##          151          152          153
## 2.571141e-04 5.921398e-03 7.284416e-04
```

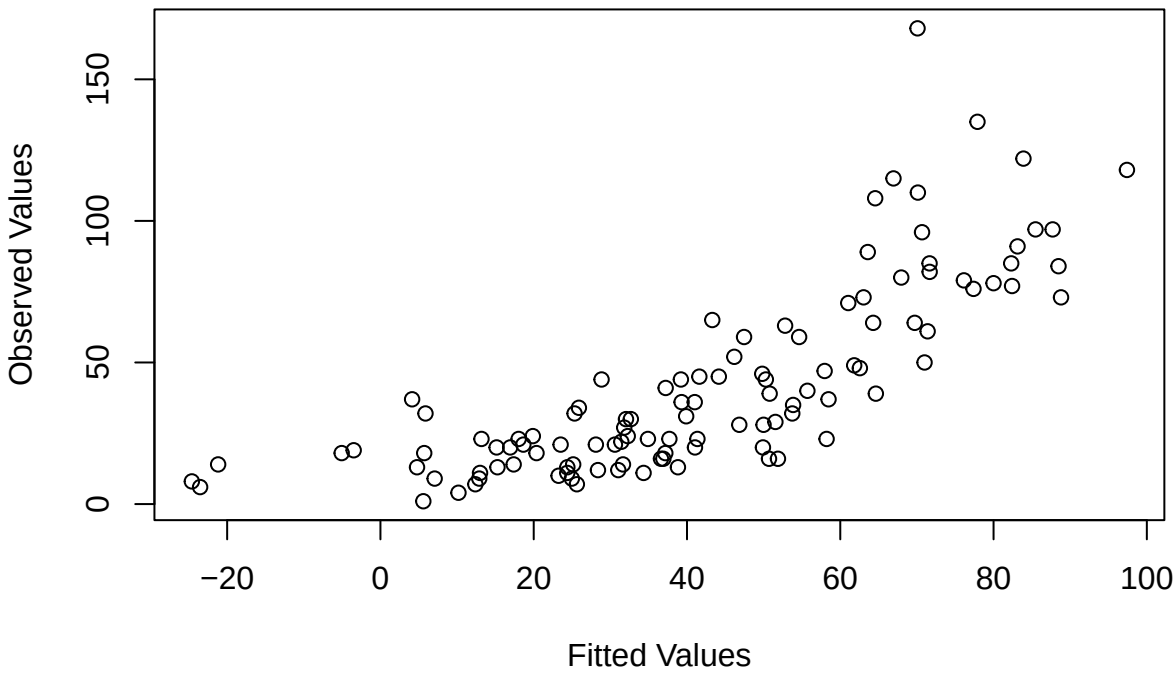
```
plot(lm_full,which=4)
```




```
hist( x = residuals(lm_full),   # data are the residuals
      xlab = "Value of residual",      # x-axis label
      main = "",                       # no title
      breaks = 20                      # lots of breaks
 )
```




```
yhat.2 <- fitted.values( object = lm_full)
 plot( x = yhat.2,
       y = airquality$Ozone,
       xlab = "Fitted Values",
       ylab = "Observed Values"
 )
```


Observed Values
0
50
100
150
−20
0
20
40
60
80
100
Fitted Values